Prediction of cancer mutation states using multiple data modalities reveals the utility and consistency of gene expression and DNA methylation

Jake Crawford[1], Brock C. Christensen[2], Maria Chikina[3], Casey S. Greene[4,5]

[1]Genomics and Computational Biology (GCB) Graduate Group, Perelman School of Medicine, University of Pennsylvania
[2]Department of Epidemiology, Geisel School of Medicine, Dartmouth College
[3]Department of Computational and Systems Biology, School of Medicine, University of Pittsburgh
[4]Department of Biochemistry and Molecular Genetics, University of Colorado School of Medicine
[5]Center for Health AI, University of Colorado School of Medicine

Background

Research questions:

- Which -omics type captures the functional signatures of cancer mutations most effectively? Is this dependent on the gene(s) that are mutated?
- Does combining multiple -omics types improve detection?

Framing as a prediction problem:

We want to predict cancer mutation presence or absence using -omics data in the TCGA Pan-Cancer Atlas: gene expression, DNA methylation, reverse phase protein array (RPPA), microRNA, somatic mutational signatures.

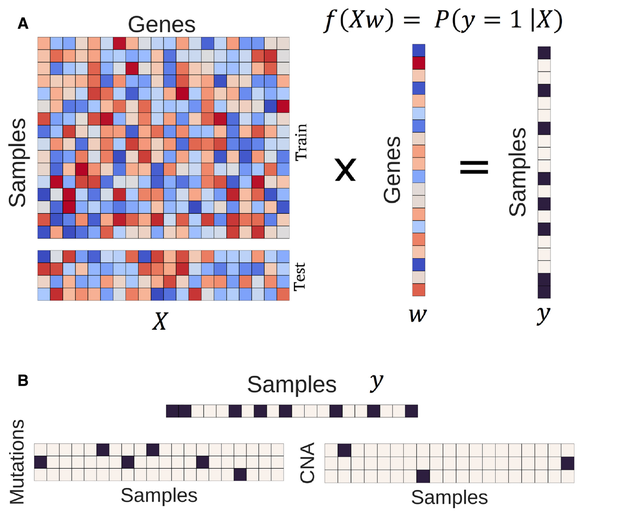
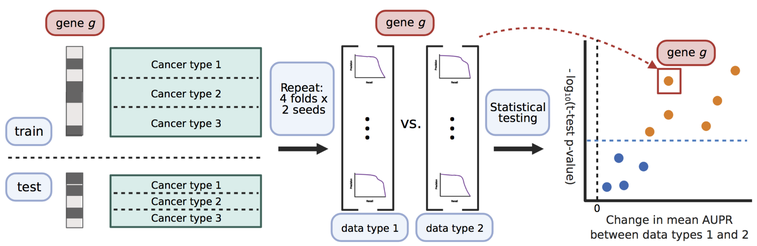


Figure from Way et al. 2018[1]

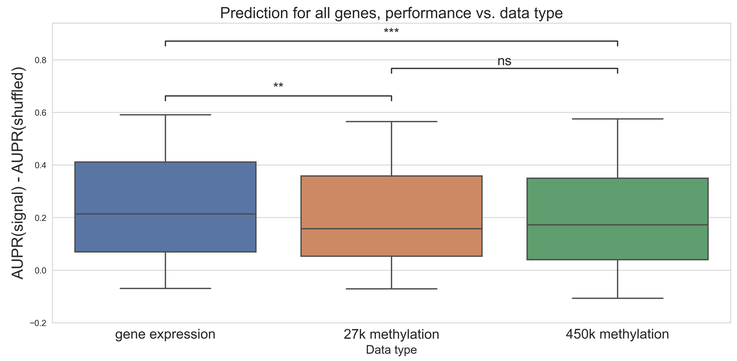
Approach

- Cancer gene set from Vogelstein et al. 2013[2], ∼85 cancer-related genes
- Elastic net logistic regression
- 2 replicates (random seeds) x 4-fold CV, stratified by cancer type
- Compare classifiers against baseline with permuted labels, and compare directly between data types

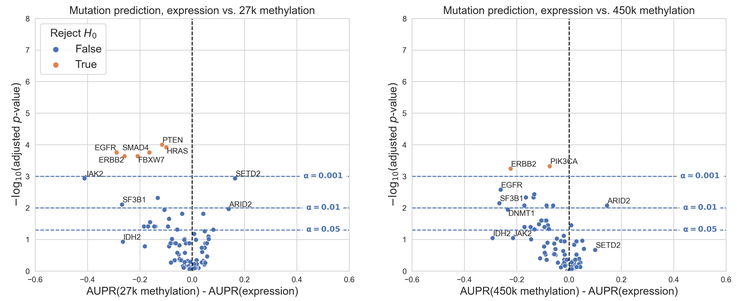


Results

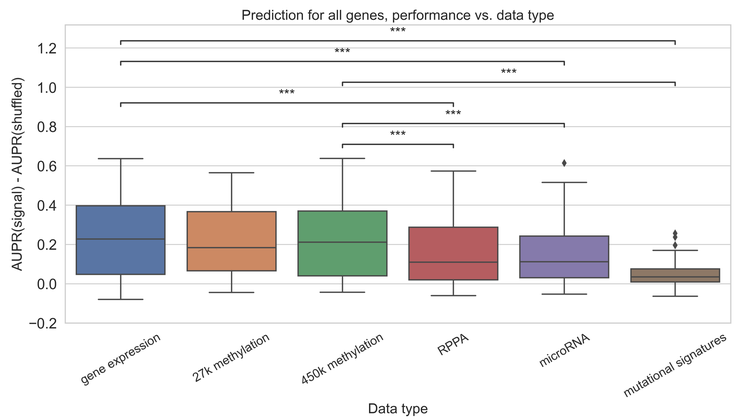
On aggregate over the Vogelstein et al. gene set, gene expression is a slightly more effective predictor than the methylation arrays (Illumina 27K/450K merged and Illumina 450K).



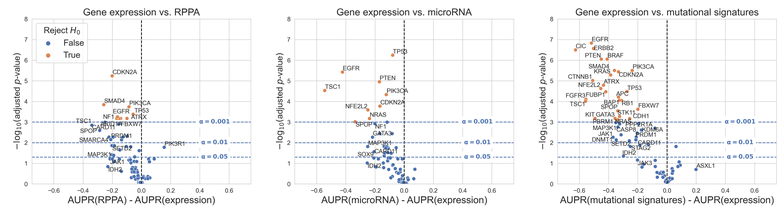
Looking at performance for individual genes, however, most genes do not significantly differ between data types (data points around origin).



When we compare all data types using all cancer genes, the expression and DNA methylation datasets significantly outperform the remaining data types.

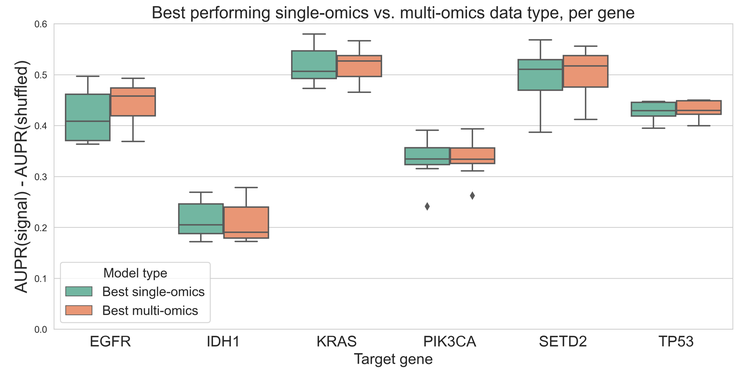


For the remaining data types, on the individual gene level, gene expression generally provides better performance (genes/points in the top left).



We also built multi-omics models by concatenating combinations of the expression and methylation datasets. For each data type, we used the top 5000 principal components as predictive features.

Using six pan-cancer driver genes as targets, none of the multi-omics models significantly outperformed the best-performing single-omics model.



We anticipate that these results will be useful in study design: ***gene expression*** and ***DNA methylation*** are ∼equally effective as a functional readout.

Relevant Links

Data and code availability:

https://github.com/greenelab/mpmp

Draft of manuscript (currently in-progress using Manubot[3]):

https://greenelab.github.io/mpmp-manuscript/

Link to this poster:

http://jjc2718.github.io/ismb_2021_poster.pdf

References

[1] G. P. Way, F. Sanchez-Vega, K. La, J. Armenia, W. K. Chatila, A. Luna, C. Sander, A. D. Cherniack, M. Mina, G. Ciriello, *et al.*, "Machine learning detects pan-cancer Ras pathway activation in The Cancer Genome Atlas," *Cell Reports*, vol. 23, no. 1, pp. 172–180, 2018.

[2] B. Vogelstein, N. Papadopoulos, V. E. Velculescu, S. Zhou, L. A. Diaz, and K. W. Kinzler, "Cancer genome landscapes," *Science*, vol. 339, no. 6127, pp. 1546–1558, 2013.

[3] D. S. Himmelstein, V. Rubinetti, D. R. Slochower, D. Hu, V. S. Malladi, C. S. Greene, and A. Gitter, "Open collaborative writing with Manubot," *PLoS Computational Biology*, vol. 15, no. 6, p. e1007128, 2019.